An Open-Source Interface to the Canadian Surface Prediction Archive

Martin Gauch,[1] James Bai,[1] Juliane Mai,[2] and Jimmy Lin[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of Civil and Environmental Engineering, University of Waterloo

ABSTRACT

Data-intensive research and decision-making continue to gain adoption across diverse organizations. As researchers and practitioners increasingly rely on analyzing large data products to both answer scientific questions and for operational needs, data acquisition and pre-processing become critical tasks. For environmental science, the Canadian Surface Prediction Archive (CaSPAr) facilitates easy access to custom subsets of numerical weather predictions. We demonstrate a new open-source interface for CaSPAr that provides easy-to-use map-based querying capabilities and automates data ingestion into the CaSPAr batch processing server.

ACM Reference Format:
Martin Gauch, James Bai, Juliane Mai, and Jimmy Lin. 2020. An Open-Source Interface to the Canadian Surface Prediction Archive. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398626

1 INTRODUCTION

Researchers and practitioners in the environmental sciences use numerical weather predictions to publish short-term weather forecasts, analyze historic climate events, and model climate change. It is, however, difficult to manipulate these data products. Perhaps the biggest challenge is that they are often very large, encompassing larger temporal and spatial domains than the individual user requires. For example, the researcher might only be interested in a few variables around Lake Erie during the winter of 2018, but the relevant source contains data spanning several decades across large swaths of the entire Great Lakes region. This scenario exemplifies what we refer to as the *subsetting problem*: typically, a researcher solves this by downloading the *entire* data product (or a large fraction thereof, depending on the collection's physical organization) and then manually extracting only the needed spatial and temporal domains. Needless to say, this is a slow and laborious process. Compounding this challenge is the fact that data products are scattered across various websites, organized differently, and use different formats. As the popular saying goes, "80% of data science is data cleaning"—this is certainly true in our domain: researchers spend most of their time gathering and pre-processing data, rather than focusing on their actual research or operational questions.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08.
https://doi.org/10.1145/3383583.3398626

In 2017, Mai et al. [2] introduced the Canadian Surface Prediction Archive (CaSPAr, caspar-data.ca) as a web service to facilitate access to weather predictions provided by Environment and Climate Change Canada (ECCC). CaSPAr is a web application with a Google Maps-like interface where researchers can request arbitrary subsets of ECCC data products, exactly matching the scenario described above: Users can choose from a list of available data products. After selecting one, the map outlines the product's geographic domain; users can either indicate their regions of interest directly on the map or upload a shapefile or GeoJSON specification. To complete a request, users also indicate the desired date ranges and the variables of interest. Upon submission of a request to the CaSPAr batch processing server, automated scripts process the selected data product and extract the desired subset, storing the results in NetCDF files that follow the CF-1.6 standard. After this processing completes, the user receives an email with a download link through the Globus large file transfer service (globus.org).

CaSPAr currently contains five operational numerical weather forecast products, four operational analyses, and one reanalysis product. These products are issued up to eight times a day and have forecast horizons of up to 16 days. Their spatial resolutions range between 2.5 km and 50 km. The CaSPAr documentation at github.com/julemai/CaSPAr/wiki provides further details about the products. As of June 2020, the archive contains 444 TB of data—a number that increases by 368 GB of newly created data each night. This service greatly alleviates researchers' access to data—currently (June 2020), CaSPAr has 95 active users.

2 A NEW INTERFACE

Although CaSPAr has been operational since 2017, several challenges have emerged. For example, the current design made it difficult to accommodate the dynamic nature of the incoming data. On a regular basis the resolution of products is increased, new variables and forecast horizons are added, and more issues are produced. Furthermore, the current CaSPAr interface uses proprietary, closed-source software by Esri to display data products on a map and to manage the selection of irregularly-shaped spatial domains. Although the Esri software seemed to be the best choice for this application when CaSPAr was built, capturing the dynamic nature of products has turned out to be very labour-intensive: Every change in the products needs to be manually ingested in the interface, which, in the long term, is not sustainable.

This demonstration describes our efforts in building a new interface for CaSPAr entirely based on open-source software and libraries, that reduces manual maintenance work to a minimum. Figure 1 shows a screenshot of the new web UI. Our new interface automates the nightly data ingestion through a REST API that allows for incremental updates to the data products. Each day, the

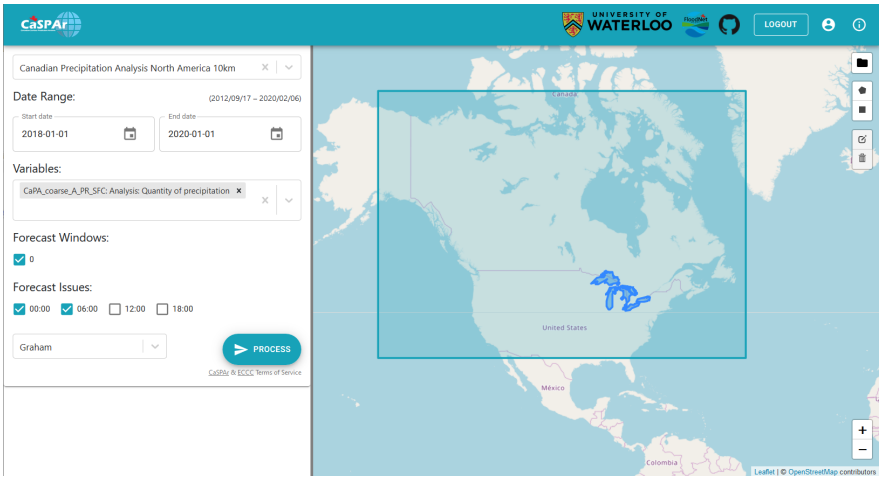

Figure 1: Screenshot of the new CaSPAr web interface. The user has uploaded a shapefile of the Great Lakes region.

batch processing server receives ECCC's newly issued weather predictions as FST files, analyzes their contents, converts them to NetCDF files compliant with CF-1.6, adds additional metadata, and subsequently registers the changed or updated properties of the data products in the metadata database of the new interface.

CaSPAr consists of two major components: the new, lightweight interface, implemented in JavaScript (React) and Python (Flask) docker containers, and an existing batch processing service, implemented in Python, bash, and Fortran, which remains unchanged. Figure 2 shows a schematic architectural diagram of the interaction between the user's browser, the interface server, and the batch processing server. Whenever a user opens the CaSPAr website, the browser loads the React site, which requests the available products' properties from the Flask application, where they are stored in a PostgreSQL database.

Once a user sends a request for a product, the Flask application forwards it via `scp` as a JSON file to the batch processing server, which is hosted in a Compute Canada high-performance computing environment. The server receives the request and schedules a job to subset the specified product. Once this job completes, the processing service notifies the user via email that their data is ready to download via Globus file transfer and reports statistics on execution status, job duration, and required storage to the interface.

3 CONCLUSION

As data-intensive science fosters its status as the "fourth paradigm" of research [1], easy access to data products becomes key to successful scientific work. Automated, reusable, and open-source solutions such as the CaSPAr web interface alleviate this access and leave researchers more time for the work they actually care about.

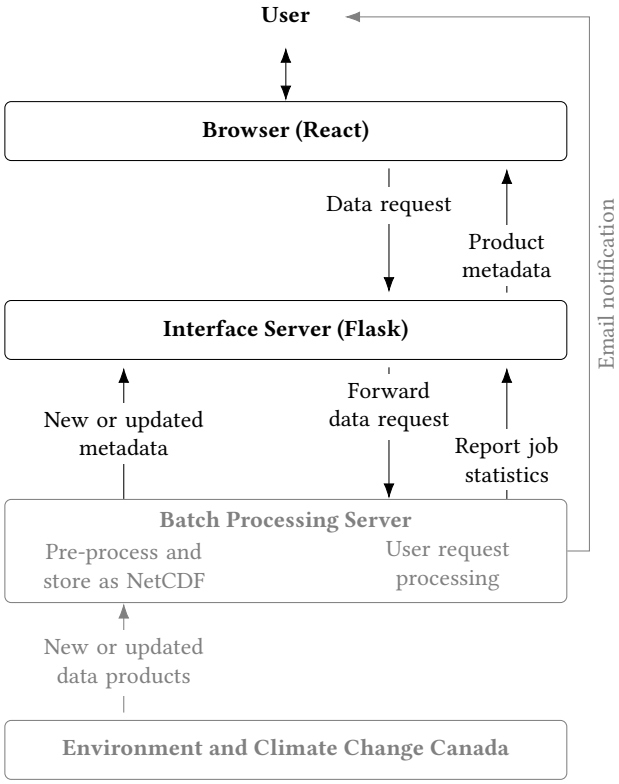

Figure 2: Schematic diagram of the CaSPAr architecture and control flow. Existing components are colored in gray.

ACKNOWLEDGMENTS

The additions to CaSPAr described here were supported by the Canada First Research Excellence Fund (Global Water Futures). CaSPAr is funded by NSERC's FloodNet and Environment and Climate Change Canada. Additional support for computational resources were provided by Compute Ontario and Compute Canada.

REFERENCES

[1] T. Hey, S. Tansley, and K. M. Tolle (Eds.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research.

[2] J. Mai, K. C. Kornelsen, B. A. Tolson, V. Fortin, N. Gasset, D. Bouhemhem, D. Schäfer, M. Leahy, F. Anctil, and P. Coulibaly. 2019. The Canadian Surface Prediction Archive (CaSPAr): A Platform to Enhance Environmental Modeling in Canada and Globally. *Bulletin of the American Meteorological Society* 101, 3 (2019), E341–E356.